

Real-time Hypoglycemia Risk Management: A Calibrated and Interpretable HMM Approach for Continuous Glucose Monitoring

Arumugam P.¹, Sornalatha M. E.², Sakthi Malaviga S. R.³

¹ Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627012, Tamil Nadu, India.
Email: sixfacemsu@msuniv.ac.in

² Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627012, Tamil Nadu, India.
Email: swarnalathapramod@gmail.com

³ Manonmaniam Sundaranar University, Abishekapatti, Tirunelveli – 627012, Tamil Nadu, India.
Email: sakthimalaviga33521@gmail.com

Abstract: Objective: This study enhances clinical interpretability and calibration of real-time 60-minute hypoglycemia risk from Continuous Glucose Monitoring. It reframes Hidden Markov Model risk as a posterior-weighted probability of glucose ≤ 70 mg/dL, compared with logistic regression and a stacked ensemble. Methods: We developed an interpretable HMM P70 model for comparison of its performance with regularized logistic regression and a stacked ensemble. Evaluation, using a multi-patient CGM cohort, discrimination, calibration and net benefit Results: HMM P70 showed significantly improved calibration and lower Brier scores. However, logistic regression and ensemble models had improved discrimination, net clinical benefit, and earlier alert. The best combined performance was provided by the ensemble model. Conclusion: HMM P70 is a well-calibrated, interpretable real-time risk signal. It is complementary to more discriminative logistic regression models, which will help improve understanding and safe deployment of hypoglycemia risk tools.

Keywords: Continuous Glucose Monitoring, Real-time risk management, Clinical interpretability, Stacked Ensemble, Diabetes management

1. Introduction

CGM is a fundamental part of type 1 and 2 diabetes management compared to the traditional one that does not provide sufficient knowledge on the temporal dynamics of diabetes risk assessment ([5]). The traditional machine learning classifiers are not sensitive to the time dynamics inherent in the diabetes risk assessment ([1]). Hidden Markov Models can be inferred quickly and interpreted to latent states, but simple state means can lead to risk predictions that are not well calibrated. This study redefines HMM risk as the probability of the glucose level being a rate at the risk of being ≤ 70 mg/dL when weighted by the posterior probability, and compares this risk measure with a logistic regression model ([2]) ([6]). An accurate blood glucose prediction is essential in making an informed decision about insulin, diet, and physical activity, which results in improved quality of life, as well as other minimized health problems ([10]). Machine learning, and especially HMMs, are being identified as having potential in predicting clinical measurement sequences into useful knowledge to characterize risk and progression of diseases ([13]). It is especially important in Type 1 diabetes when consistency of glucose levels will be required to avoid long-term complications ([14]). The dynamic properties of glucose, which are affected by diet, activities, and medication conditions, necessitate an advanced modelling of effective forecasting of hypoglycemia ([7]). The detection of early hypoglycemia is a very important issue because it is both potentially life-threatening and asymptomatic, which is why it is essential to have very powerful models that can offer well-defined actionable risks that can be assessed on the



self, and negative glycemic events can be prevented ([6]) ([17]). The conventional approaches to predicting blood glucose levels, including ARIMA and LSTM models, have been significantly used to forecast different glucose levels, but they might not be able to represent the non-linear processes of glucose regulation in real-time ([11]). Therefore, it necessitates a more confident prognostic model, which will allow the identification of the high-risk group of persons and predict their future glycemic conditions. This is beneficial in targeted interventions and better patient outcomes ([13]). Also, this form of framework facilitates a holistic analysis of the model performance on different platforms (e.g., statistical discrimination, calibration, and decision-analytic utility), which is essential in evaluating predictive AI models in healthcare environments ([3]). This scientific assessment tool is applied to guarantee that the models are statistically sound, in addition to being clinically actionable; hence the issue of reliability and applicability in practical settings is addressed in this writing ([12]). Honest reporting practices and ethical conduct support the scientific methodology applied. The compliance to data preprocessing, hyperparameters optimization, and healthy model testing is necessary to produce quality machine learning systems within the medical field ([3]). This is an intensive process, starting with the preparation of data and concluding with the validation of the data and the fact that development of predictive, reliable models of hypoglycemia is a complex and multifaceted issue ([4]) ([8]). This thorough validation, with a focus on clinical utility and interpretability, prevails in the successful application of machine learning applications to endocrinology and managing diabetes, where the priority is not so much on theoretical quantitative features but on actual patient outcome ([9]) ([15]). The novelty of this work is the introduction of a specific state-space structure of the prediction of real-time, clinically interpretable hypoglycemia based on Hidden Markov Models. Earlier versions of HMM implementations have difficulty in risk predictions that are not well calibrated. This is directly approached by recasting the HMM risk as a posterior-weighted entity of glucose levels being 70mg/dL or less (P70 score). This novel change greatly enhances the calibration, resulting in a better and clinically interpretable signal. We introduce a complete working framework of real-time management of hypoglycemia risks that are benchmarked with a regularized logistic regression model and a stacked ensemble.

2. Methods

2.1 Data Sources and Population of the study.

Two continuous glucose data sets, one of which is an internal development cohort secondary data as used by([4]), and the other is an external validation cohort (secondary data as used by ([16])). The internal cohort sample consisted of 6 Type 1 diabetes study participants (873,702 CGM timepoints; median 146,197 /patient), and the external sample consisted of 11 patients (80,329 CGM timepoints; median 2,749/patient) in a different clinical unit. The people needed to possess no less than 72 hours of data from CGM and no less than 1 episode of hypoglycemia. Those sessions with more than 30 per cent gaps in CGM data or sensor warming up / calibration issues are excluded. Preprocessing of CGM values was accomplished to 5 5-minute grid, which removed artefacts. The result indicates that the action of Hypoglycemia was largely varied, with an internal median of 0.0005 and an external median of 0.0105, which is typical of dissimilar clinical states.

2.2 Outcome Definition

The primary prediction was regarding instant hypoglycemia within a 60-minute period. Any CGM below 70mg/dL for a duration beyond 15 minutes at the subsequent 60 minutes[t,t+60] is regarded to be positive. In the case of alert evaluation, a true-positive alert was seen to have occurred 60 minutes after the initial occurrence of the event with 60 60-minute refractory period. This will ensure unnecessary alerts are eradicated.

2.3 Predictive Models

We have evaluated three model classes which forecast 60-minute risk of hypoglycemia:

A Gaussian-emission HMM was trained on the input features of [glucose, 5-minute delta], and the multivariate sequences were trained. The parameters were fitted using the expectation-maximisation algorithm on the internal cohort. To enable clinically viable dynamics of CGM, our HMM was trained to include four hidden states

(i.e. Stable, Declining, Rapidly Declining, High Risk) to enable interpretation of risks in a transparent manner and provide a straightforward mapping to physiological glucose levels. Parameters in emission distributions of each state were parameterized with CGM values and time properties, and transition probabilities were learnt through maximum likelihood. In order to be usable in real-time, filtering posteriors were used. To minimize variation in the prediction of the risk in the noted rates of occurrences, isotonic regression calibration was applied, which increased reliability.

A penalized L2 logistic regression model was trained using the internal cohort and such features as current CGM, deltas, short-term variability and temporal covariates. The cross-validation was used to select the hyperparameters.

The HMM and logistic regression risk scores were stacked, enabling a shallow logistic regression meta-learner together with stacking on the patient level to remove optimistic bias. All the models were trained and only adjusted in the internal cohort; they were to be independently validated in the external cohort.

2.4 Model Training, Calibration and Risk Scoring.

Models were trained on internal cohort, temporally ordered data with splits to prevent leakage. Repeated cross-validation was used for hyperparameter tuning. Raw scores were put in scales to estimate 60-minute risk, and where it was essential, post-hoc calibration was implemented. Per-timepoint risk scores were used as a basis of analysis.

2.5 Discrimination and Overall Accuracy

Model discrimination was assessed using AUC and AUPRC. Overall probabilistic accuracy was quantified via the Brier score. 95% confidence intervals were computed using patient-level cluster bootstrapping.

2.6 Calibration

Calibration was evaluated by comparing mean predicted probability in risk bins against observed hypoglycemia rates. Calibration curves and Brier scores summarized performance, with internal/external curves compared for transportability.

2.7 Decision-Analytic Evaluation

Decision curve analysis quantified the trade-off between false positives and safety margin. Net benefit was calculated across risk thresholds and compared against "treat-all" and "treat-none" strategies, with 95% confidence intervals from cluster bootstrapping.

2.8 Alerting Configurations and Lead-time Analysis

Alerting thresholds are aimed at approximately one alert per patient-day in the internal cohort. Lead times for true-positive alerts (hypoglycemia within 60 minutes) were summarized by mean and empirical quantiles (P10, P25, P50, P75, P90) to understand warning times under a fixed alert burden.

2.9 Statistical Analysis

All analyses were timepoint-level, with patient clustering addressed via cluster-bootstrap for 95% confidence intervals on AUC, Brier score, net benefit, and lead-time quantiles. Model development and evaluation were implemented in Python using reproducible pipelines.

3. Results

This section presents the results concerning discrimination, calibration, and decision-analytic performance of the evaluated models across internal and external cohorts. From Table 1, the logistic regression model generally outperformed HMM in AUROC and AUPRC, indicating superior discrimination, while HMM P70 consistently achieved lower Brier scores, suggesting better calibration.

Table 1: Discrimination, calibration, and decision-analytic performance by cohort and model. AUC and Brier score summarize discrimination and overall accuracy.

Cohort	Model	AUC	Brier	Net Benefit (1%)	Net Benefit (5%)	Net Benefit (10%)
Internal	HMM	0.853	0.032	0.0278 (0.0161, 0.0388)	0.0135 (0.0050, 0.0231)	0.0000 (0.0000, 0.0000)
Internal	Logistic	0.880	0.055	0.0149 (0.0102, 0.0211)	0.0101 (0.0062, 0.0148)	0.0062 (0.0035, 0.0088)
Internal	Ensemble	0.841	0.032	0.0251 (0.0127, 0.0404)	0.0145 (0.0060, 0.0233)	0.0021 (0.0016, 0.0028)
External	HMM	0.812	0.022	0.0030 (-0.0003, 0.0058)	-0.0110 (-0.0153, -0.0067)	-0.0290 (-0.0415, -0.0174)
External	Logistic	0.865	0.008	0.0017 (-0.0023, 0.0045)	0.0010 (-0.0002, 0.0021)	-0.0003 (-0.0006, -0.0001)
External	Ensemble	0.825	0.008	-0.0023 (-0.0060, 0.0011)	-0.0086 (-0.0128, -0.0050)	-0.0001 (-0.0003, 0.0001)

From Table 2, the lead-time performance obtained is so good that the Ensemble model can deliver clinically significant early warning of hypoglycemia in most of the true-positive cases, which can be 15 to 45 minutes with a median around about 30 minutes, and this is the time interval that is needed to administer carbohydrates and a change in insulin dose. Even the small percentage of near-immediate occurrences has to be considered, which implies that the safety layers and conservative settings have to be put into place in a high-risk situation. Overall, the allocation and fair operating thresholds and alert hygiene should be provided with sensible proactive prevention and a moderate alert load.

Table 2: Distribution by model and cohort Lead-time distribution (minutes) of true-positive alerts at a budget of one alert per patient-day.

Model	Cohort	Mean	P10	P25	P50	P75	P90
Ensemble	Internal	19	5	5	10	30	37
	External	5	5	5	5	5	5
HMM	Internal	19	5	5	5	30	55
	External	26	8	13	20	40	50
Logistic	Internal	20	5	5	15	25	45
	External	-	-	-	-	-	-

At a strict budget of one alert per patient-day, true-positive alerts generally occurred tens of minutes before hypoglycemia, with median lead times between approximately 5 and 20 minutes. For instance, in the external cohort, the HMM produced longer median lead times of approximately 20 minutes (90th percentile ~50 minutes). In contrast, the logistic regression model yielded no true-positive alerts at this constrained operating point, highlighting the difficulty of achieving early warning with very low alert rates.

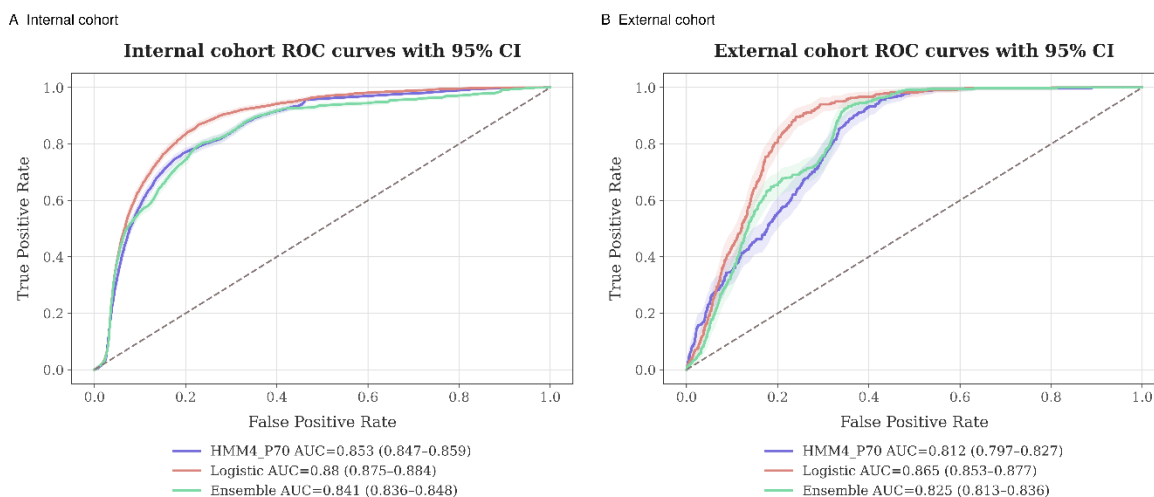


Fig. 1. Receiver operating characteristic curves for the HMM, Logistic, and Ensemble models in the internal and external cohorts.

As shown in Figure 1, all three models achieved high discrimination in both cohorts (AUC 0.84–0.88 internally and 0.81–0.86 externally), indicating a strong ability to rank-order timepoints by near-term hypoglycemia risk despite differences in case mix.

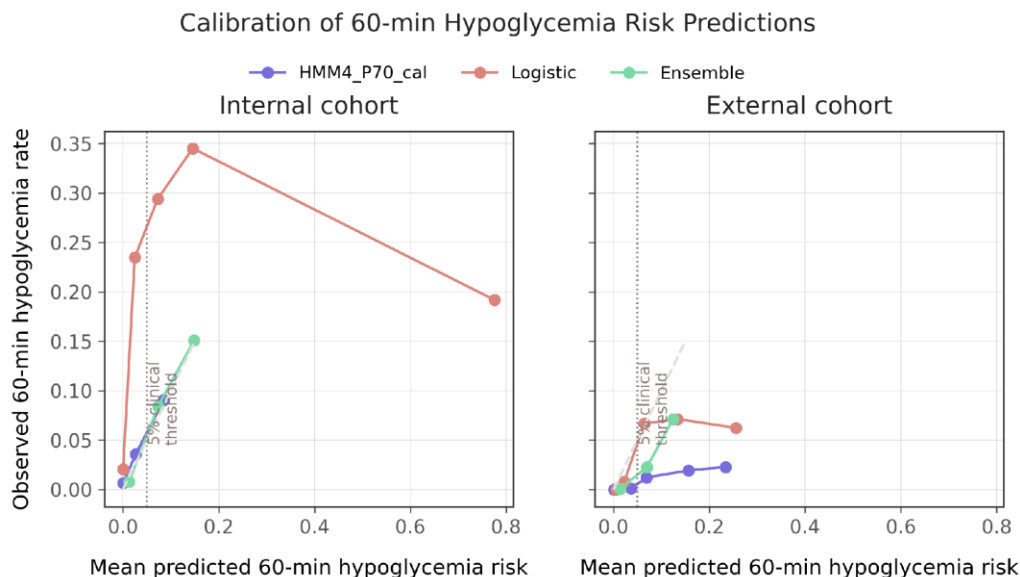


Fig.2. Calibration of 60-minute hypoglycemia risk predictions.

Calibration plots for the HMM, regularized logistic regression, and ensemble models in the internal (A) and external (B) cohorts.

Figure 2 graphically demonstrates the HMM's superior calibration, especially for the internal cohort and its relatively maintained performance externally, directly supporting the choice of the HMM P70 for its reliable risk estimates

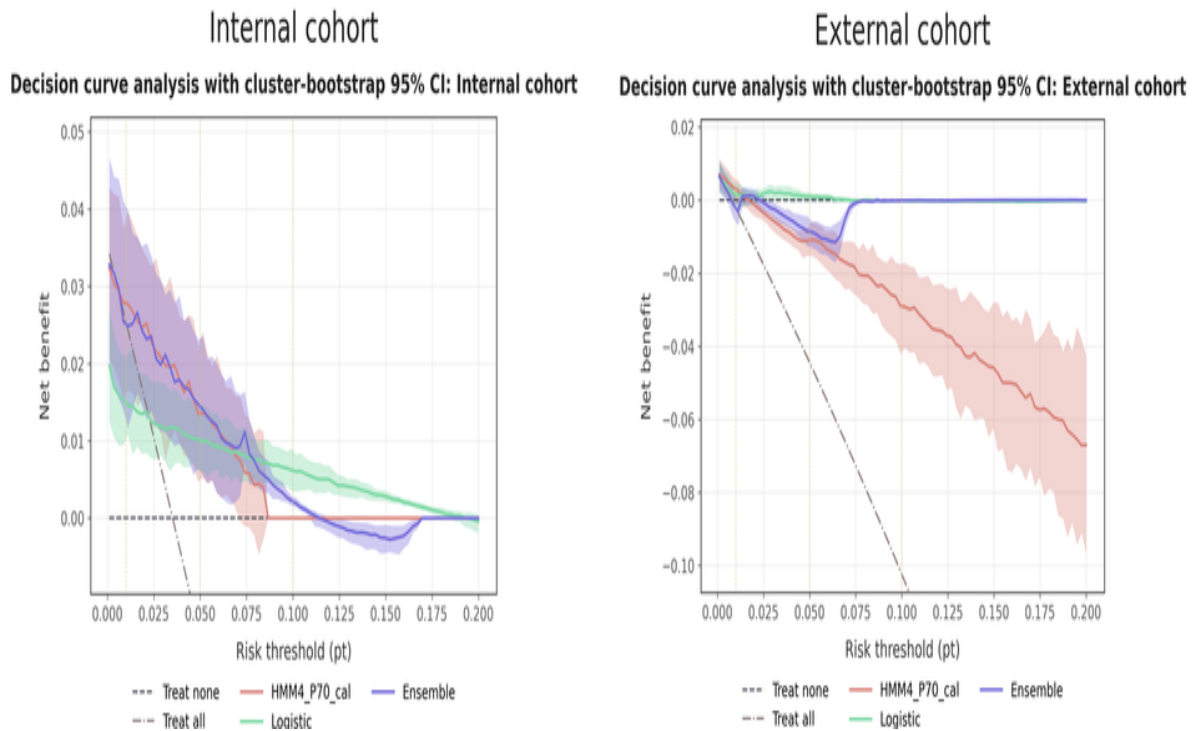


Fig .3. Decision curve analysis comparing internal and external cohorts for 60-minute hypoglycemia prediction.

As shown in Figure 3 in the internal cohort, the HMM- and ensemble-based models achieve higher net benefit than both treat-all and treat-none at clinically relevant low thresholds, with bootstrap confidence intervals consistently above zero. In the external cohort, net benefit is markedly attenuated, and confidence intervals frequently cross zero, indicating that apparent internal gains in clinical utility do not generalize fully to the external population despite preserved discrimination.

To quantify the trade-off between false positives and safety margin, decision curve analysis with patient-level cluster bootstrap confidence intervals was performed in both internal and external cohorts. In the internal cohort, at a 1% decision threshold, the ensemble model (net benefit of 0.0251) and HMM (0.0278) consistently provided higher net benefit across thresholds up to 5–10% compared to logistic regression (0.0149), reflecting a better balance between avoided hypoglycemic events and false alarms.

However, in the external cohort, net benefits were smaller and often near zero. The logistic model showed marginal positive benefit (0.0010 at 5% threshold) while HMM and ensemble occasionally yielded negative net benefits (e.g., -0.0110 for HMM). This suggests that conservative thresholds and better-calibrated models, like logistic regression, may be preferable in new populations if minimizing unnecessary interventions is paramount. Overall, decision curves highlight that lower thresholds (1–5%) favour safety, with HMM and ensemble yielding more net benefit internally, but at higher thresholds, differences narrow, and externally, same models can become net harmful. Table 1 reports the corresponding performance outcomes.

In decision-curve analysis (Figure 3), the HMM and ensemble models provided clear net benefit over treat-all and treat-none strategies in the internal cohort at low decision thresholds. In contrast, in the external cohort, net benefit was small and often indistinguishable from zero across clinically relevant thresholds. The external cohort, decision

curves showed minimal or negative net benefit at clinically relevant thresholds, likely reflecting differences in outcome prevalence and calibration drift relative to the internal development cohort, rather than a complete loss of discrimination.

4. Discussion

A 4-state Hidden Markov Model was chosen for flexibility, interpretability, and robustness. Its four latent states (stable, rising, falling, pre-hypoglycemic) align with distinct glucose regimes; fewer states merged patterns, while more were difficult to interpret.

Statistical comparisons (2–6 states) on the internal cohort showed fit improved up to 4 states (AIC/BIC, log-likelihood). Beyond this, gains were marginal; some models overfit. The 4-state HMM provided stable, reproducible state sequences and clear glucose distributions, vital for clinical understanding and real-time risk communication. This proved to be a primary model that was able to achieve clinically relevant dynamics without becoming uninterpretable and irrelevant.

Our study demonstrated a high degree of discrimination of the probabilistic models generated from CGM data in predicting 60-minute risk of hypoglycemia and the improvements provided net benefits of clinical significance, although these improvements were small under dataset shift. All three models had good AUC with a low Brier score for the internal cohort. HMM and ensemble methods showed a beneficial net effect at a low decision threshold that can be indicative of an early alarm for hypoglycemia.

In an external cohort, the discrimination was still high, and the calibration was getting worse; therefore, there was little or no net benefit. This therefore infers that thresholds in terms of nominal risk may not necessarily translate to clinical benefit if not recalibrated adequately. It again calls back to the fact that AUC alone can not be used for a safety-critical deployment, and more calibration and explicit decision analytic performance evaluation across cohorts is needed.

The following are the Limitations, which are a small cohort size, a simple set of features, 60-minute fixed prediction horizon and no adaptive online recalibration process. So, in future efforts, the focus should be on models trained on much wider and more diverse populations and on the use of much richer context data, and looking for exciting recalibration techniques.

5. Conclusion

In humour, the rethinking of HMM risk as a posterior-weighted probability of glucose $\leq 70\text{mg/dL}$ works wonders in the calibration. High discrimination, low Brier scores, and positive net benefit at low decision thresholds in the internal cohort using a model with an interpretable four-state HMM, logistic regression, and a stacked ensemble. This included supporting strategies for early warning to reduce missed events without excessive alert burden. However, in an external cohort, these strengths proved fragile. Discrimination remained high, but calibration degraded, and decision-curve analyses showed little or no net benefit. This indicates AUC alone overstates deployability; calibration and decision-analytic performance require explicit evaluation across settings. These results illustrate HMM's promise and limitations for CGM-based decision support. Using the logistic regression model as the primary alerting signal with HMM P70 for calibration, interpretability, and ensembling could be endorsed as a reliable deployment strategy. Also, future work focusing on ideal validation involving larger, diverse populations, incorporating richer contextual data (e.g., insulin dosing, meals), and robust recalibration strategies is recommended for safe clinical use. HMM P70 provides a calibrated, interpretable risk signal that complements logistic regression models and is suitable for real-time deployment.

Recommendations and Next Steps:

- Combine strengths: Investigate calibrated model stacking or decision rules for alert triggers.

- Prospective validation: Assess alerting performance prospectively, externally, with protocolized alert budgets and clinician-relevant endpoints.
- Calibration in the large: Report slope/intercept and consider lightweight recalibration.
- Broaden sensitivity analyses: Vary horizon/refractory windows, event definitions; evaluate subgroup performance.

Declaration Of Competing Interest

The authors declare that they have no known competing financial interests or personal relationships that could influence the work reported in this article.

Funding

This study was supported by the Program of Manonmaniam Sundaranar University. Sornalatha M E was supported by an M.G. Ramachandran Centenary Fellowship.

5. Acknowledgement

The second author gratefully acknowledges financial support from Manonmaniam Sundaranar University through the M.G. Ramachandran Centenary Fellowship. (MSU/R/RES/R2).

REFERENCES

1. Mohamad Al Ahdab, Milan Papež, Torben Knudsen, Tinna Björk Aradóttir, Signe Schmidt, Kirsten Nørgaard, and John Leth. 2023. Stochastic Modeling and Inference for Type 2 Diabetes. (07 2023). doi:10.36227/techrxiv.23743938.v1
2. Hatim Butt, Ikramullah Khosa, and Muhammad Aksam Iftikhar. 2023. Feature Transformation for Efficient Blood Glucose Prediction in Type 1 Diabetes Mellitus Patients. *Diagnostics* 13 (01 2023), 340–340. Issue 3. doi:10.3390/diagnostics13030340
3. Ben Van Calster, Gary S. Collins, Andrew J. Vickers, Laure Wynants, Kathleen F. Kerr, Lasai Barreñada, Gaël Varoquaux, Karandeep Singh, Karel Moons, Tina Hernandez-Boussard, D. Timmerman, David J. McLernon, Maarten van Smeden, and Ewout W. Steyerberg. 2024. Performance evaluation of predictive AI models to support medical decisions: Overview and guidance. *arXiv (Cornell University)* (12 2024). doi:10.48550/arxiv.2412.10288
4. Danilo F. Carvalho, Uzay Kaymak, Pieter Van Gorp, and N.A.W. van Riel. 2022. A Markov model for inferring event types on diabetes patients data. *Healthcare Analytics* 2 (01 2022), 100024–100024. doi:10.1016/j.health.2022.100024
5. Simon Lebech Cichosz, Stine Hangaard, Thomas Kronborg, Peter Vestergaard, and Morten Hasselstrøm Jensen. 2025. From Data to Insights: A Tool for Comprehensive Quantification of Continuous Glucose Monitoring (QoCGM). *medRxiv (Cold Spring Harbor Laboratory)* (01 2025). doi:10.1101/2025.01.01.25319870
6. Beyza Cinar, Jennifer Daniel Onwuchekwa, and Maria Maleshkova. 2025. Deep Learning-Based Hypoglycemia Classification Across Multiple Prediction Horizons. (01 2025). doi:10.48550/ARXIV.2504.00009
7. Ebrahim Farahmand, Reza Rahimi Azghan, Nooshin Taheri Chatrudi, Eric Kim, Gautham Krishna Gudur, Edison Thomaz, Giulia Pedrielli, Pavan Turaga, and Hassan Ghasemzadeh. 2025. AttenGluco: Multimodal Transformer-Based Blood Glucose Forecasting on AI-READI Dataset. (01 2025). doi:10.48550/ARXIV.2502.09919
8. Jesper Fleischer, Troels Krarup Hansen, and Simon Lebech Cichosz. 2022. Hypoglycemia event prediction from CGM using ensemble learning. *Frontiers in Clinical Diabetes and Healthcare* 3 (12 2022). doi:10.3389/fcdhc.2022.1066744
9. Namki Hong, Heajeong Park, and Yumie Rhee. 2020. Machine Learning Applications in Endocrinology and Metabolism Research: An Overview. 71-84 pages. Issue 1. doi:10.3803/enm.2020.35.1.71
10. Heydar Khadem, Hoda Nemat, Daisy Elliott, and Mohammed Benaissa. 2023. Blood Glucose Level Time Series Forecasting: Nested Deep Ensemble Learning Lag Fusion. *Bioengineering* 10 (04 2023), 487–487. Issue 4. doi:10.3390/bioengineering10040487
11. Umaphathi Krishnamoorthy, V. Karthika, M K Mathumitha, Hitesh Panchal, Vijay Kumar S. Jatti, and Abhinav Kumar. 2023. Learned prediction of cholesterol and glucose using ARIMA and LSTM models – A comparison. *Results in Control and Optimization* 14 (12 2023), 100362–100362. doi:10.1016/j.rico.2023.100362
12. Oyebayo Ridwan Olaniran, Aliu Omotayo Sikiru, Jeza Allohibi, Abdulmajeed Atiah Alharbi, and Nada Alharbi. 2025. Hybrid Random Feature Selection and Recurrent Neural Network for Diabetes Prediction. *Mathematics* 13 (02 2025), 628–628. Issue 4. doi:10.3390/math13040628

13. Sajida Perveen, Muhammad Shahbaz, Karim Keshavjee, and Aziz Guergachi. 2019. Prognostic Modeling and Prevention of Diabetes Using Machine Learning Technique. *Scientific Reports* 9 (09 2019). Issue 1. doi:10.1038/s41598-019-49563-6
14. Sean Pikulin, Irad Yehezkel, and Robert Moskovitch. 2024. Enhanced blood glucose levels prediction with a smartwatch. *PLoS ONE* 19 (07 2024). Issue 7. doi:10.1371/journal.pone. 0307136
15. Yashpal Ramakrishnaiah, Nenad Macesic, Geoffrey I. Webb, Anton Y. Peleg, and Sonika Tyagi. 2025. EHR-ML: A Data-Driven Framework for Designing Machine Learning Applications with Electronic Health Records. *International Journal of Medical Informatics* (01 2025), 105816– 105816. doi:10.1016/j.ijmedinf.2025.105816
16. Padi Tirupati Rao and Surnam Narendra. 2024. Two layered hidden Markov model for studying type 2 diabetes. *International Journal of Systems Assurance Engineering and Management* (10 2024). doi:10. 1007/s13198-024-02491-9
17. Taiyu Zhu, Chukwuma Uduku, Kezhi Li, Pau Herrero, Nick Oliver, and Pantelis Georgiou. 2022. Enhancing self-management in type 1 diabetes with wearables and deep learning. *npj Digital Medicine* 5 (06 2022). Issue 1. doi:10.1038/s41746-022-00626-5